Comparing IR-systems

Paijmans, J.J.

Publication date:
1992

Document Version
Publisher's PDF, also known as Version of record

[redacted]

ITK RESEARCH REPORT

INSTITUTE FOR LANGUAGE TECHNOLOGY AND ARTIFICIAL INTELLIGENCE INSTITUUT VOOR TAAL- EN KENNISTECHNOLOGIE

ITK Research Report No. 39

Comparing IR-Systems: CLARIT and TOPIC

Hans Paijmans

1992 / 39

ITK
Warandelaan 2
P.O. Box 90152
5000 LE TILBURG
itk@kub.nl

August 1992

COMPARING IR-SYSTEMS: CLARIT AND TOPIC.

by[1]

Hans Paijmans

Introduction

The discipline of information retrieval is perhaps as old as the written word.With the advent of big libraries much effort was put in systems that enable people to retrieve information (or rather, documents containing that information) from those collections. Nevertheless, the tone of the IR-community is not optimistic, when it talks about the performance of current systems. The observations of Cleverdon on the subject of human indexing [Cleverdon, 1984] and the famous Blair/Maron experiment in full text retrieval [Blair/Maron, 1985], still loom darkly over all attempts to substantially alter the effectivity of information retrieval techniques. The probabilistic and vector models, that have been perfected in the last ten years, however sophisticated, cannot claim to imply real understanding of the documents and the AI-approaches are hampered by the fact that the creation and maintaining of involved knowledge representations only works in very small domains. And as Cleverdon observed, human indexing just is not consistent enough to guarantee acceptable recall and precision over sizable databases.

[1]I want to thank Dr R.C Verheijen of Digital Equipment and Drs T. van den Aker of Tilburg University for the opportunity to use the CLARIT and TOPIC systems and their help in conducting the experiments.

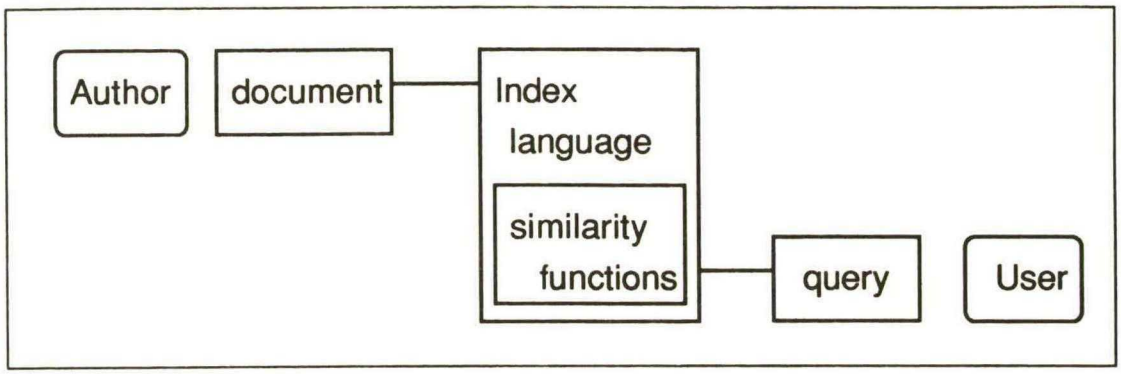

Fig. 1 The Salton model of information retrieval

But the ever increasing number of articles, documents and books in libraries and elsewhere will not disappear just because there is no unified paradigm in which to solve the retrieval problem. Therefore widely different schemes and systems have evolved to solve that problem and this fact has placed an increasing importance on strategies for testing such schemes and systems.

In this article we will apply some testing methods that were used and published by White and Griffith, to documents that were indexed by the TOPIC and CLARIT systems. Thus we hope to compare their performance relative to each other and gain some insights in certain problematic areas of IR in general.

Background.

Although there is no general accepted taxonomy for IR-systems, it is possible to recognise several models [Paijmans,1992]. A model that is widely accepted and that is used in one way or another in almost every IR system is the *boolean model.* In this model set-theory is used to handle lists of descriptors to create sets of documents that are relevant to a query. Another model that touches on our theme is the *thesaural model*, that tries to organise such descriptors in a semantic network. Other models define how to select descriptors from the document (e.g. the *probabilistic model*), or how to compare documents and queries to decide on their similarity (the *vector model*). Literature abounds with systems that experiment with one model or another, but when we look which of these models have been implemented in real life systems, the number is disappointingly small. Most systems adhere firmly to the boolean model, using an exhaustive inverted list of words as document representation (for a description of this and other concepts see below). An early and succesfull system to do this was STAIRS [IBM, 1976], which was followed both on big computers and on PC's by a multitude of epigones (e.g. Freebase or Zyindex), that really only differed in user interface.

If we look at models of retrieval systems, we notice a few parts that recur in almost every system. These parts are concisely arranged in the general IR-model as put forward by Salton [Salton/McGill, 1983] (fig.1).

Central in this model stands the Index Language (IL), into which both document and query are translated and which acts as a common reference to decide on the similarities between them and how to rank them in one order or another. An important part of this IL is a vocabulary of keywords or descriptors to describe the documents in the system, although attempts have been made to create more involved knowledge representations, such as frames.

In most IR systems documents have a place that is analogous to the record in normal database systems: i.e. a repository of data relating to a single item and we will often use the word record to mean document. This record also may act as a node to which other observations about the document are attached and in the process the document may disappear and be replaced by an abstract and bibliographic reference. Speaking about full text systems,we will use record and document as more-or-less synonym.

The parts of a document (possibly the complete text including *mark-up*s and other visual information) that are presented to the IR-system, we will call the *document surrogate*. The *document representation* (for short *docrep*) is that part of the document surrogate (or, if not a part, at least a description of it) that finds its way in the index language, thus representing the document in the system. We will see below, when we describe TOPIC, that this document representation, among other things, may well contain the complete document. The *online document* then is the document that is reported back to the user. In most cases this is a bibliographic reference or the reference and an abstract, but advanced systems may display the text of the document. Broadly speaking, the more advanced the system is, the more all three, the document surrogate, the document representation and the on-line document, will approach the original document.

Referring to the drawing in fig 1, we can see that the three main concerns of IR are:

a. translation of document and query into terms of the indexing language including
b. the modelling of the document representation and
c. supporting the user in the interaction with the index language and the knowledge contained in the document representations.

The document representation has a central role in this model, because at query time all questions will have to be matched against these docreps. Our concern in this study will be the quality of the docreps created by the two systems under consideration, given a document surrogate that contains the complete text (but without information about italics, different fonts, lay-out etc.). In particular we will analyse the document representations of the CLARIT and TOPIC systems.

Derived and assigned indexing.

There are essentially two approaches to the creation and maintenance of this document- or knowledge representation. One is to create a knowledge system in advance and assign the documents to it afterwards: *assigned indexing*. The other is to

derive the terms of the index language from the documents themselves: *derived indexing*. The manual library systems, in which books were classified according to an existing classification system, e.g. Dewey or UDC, are *assigned indexing* systems; computerized IR-systems, that extract keywords from the documents according to some weighting scheme or another, are typical for *derived indexing* systems.

We will extend the definition of assigned indexing systems to contain all systems that use terms in their docreps, that are not taken from the documents themselves, because such external terms belong to a knowledge representation outside the document.

The derived indexing systems became very popular when the computer made it easy to create an inverted list of all occurring words in a document base. In the seventies and eighties much effort was directed to techniques how to identify such words (phrases, sentences) in the inverted lists as were most efficient in retrieving particular documents. (for a discussion of both derived vs. assigned and pre-coordinate vs. post coordinate systems see: [Foskett, 1976]).

Pre- and postcoordination.

Many concepts that rightfully belong in the indexing language, may only be expressed using more than one word, e.g. *'computer programming'* or *'aluminum welding'*. In the older, manual indexing systems such compound phrases generally were recognized, kept together and entered in the system as complete terms. Such systems are called *pre-coordinate* systems. As a direct consequence of the inverted list and the ease with which sets of records could be handled using such lists, the emphasis came to lie on *post-coordinative* indexing, i.e. potential multi-word descriptors were separated in their components (or even word stems) and only these components were stored in the inverted list. At query time boolean operators such as AND, OR etc. were used to try and reconstruct such terms. Not unexpectedly much information was lost in the process and precision suffered. Therefore the *specificity* of the older, pre-coordinative systems sometimes was better than that of the newer, keyword-driven systems. Indeed we will see how Salton and Cleverdon found that such multi-word phrase indexes became overspecific and that performance dropped as compared with single-word indexes.

TOPIC and CLARIT, the two systems described here, may be considered as each belonging to one of the derived and assigned indexing systems. Also, at least CLARIT tries to combine the efficiency of the inverted file while recovering the specifity of the pre-coordinative systems: it was built to detect significant *phrases* (noun phrases). TOPIC, that superimposes a semantic hierarchy on the inverted file system, implements a kind of higher order thesaurus and in such a thesaurus the compound terms that are typical for pre-coordinate systems may also be defined.

We will argue that TOPIC (when the *topics* are used) essentially is an *assigned* indexing system, although it maintains a complete inverted file of all words in all documents. The reasons for this will be explained below. That CLARIT is a derived indexing system is easier to see, as the terms that make up the document

representation, are extracted from the original document and no other terminology is added.

If we test the two system side by side, this is not because we want to compare the two systems in a kind of consumers test. To do this would need a better understanding of both the systems and the prospective users. Also, the data used for the tests would need to be selected in a different way. We just want to do a few explorative experiments in order to gain a better understanding of the way the philopsophy of the two systems as it is displayed in the difference between the document representations, might influence performance.

Following the methodology of White and Griffith we will try to establish the performance of CLARIT and TOPIC relative to each other on three points:

1. The ability of both systems to link related documents.
2. The ability to discriminate between these linked subsets.
3. The ability to discriminate finely between individual documents.

The systems.

We have chosen the CLARIT and TOPIC systems for our tests, because the TOPIC system was selected by the new library of Tilburg University as database system for the management of the on-line contents database: a so-called *current awareness service* in which articles in journals are made accessible for retrieval immediatly after appearance (see [Roes,1992]). The system was installed on DEC equipment. This company also has supported the development of the CLARIT system through DEC's External Research Programme.

TOPIC or RUBRIC.

The first system, TOPIC by Verity Inc.[2], is the commercial offshoot of the rather well publicised experimental RUBRIC-system [McCune,a.o., 1985]. Although TOPIC is a complete system, with indexing modules, retrieval engine and a user interface for interactive querying, we will limit ourselves to the document representations and related issues.

TOPIC approaches the problem of document retrieval in two stages. In the first stage a complete inverted file is created of all occurring strings in the document. Positional information about paragraphs or particular segments of the documents, is preserved in this inverted file. Together with Boolean and proximity operators and the ability to recognise fields in the document, this puts TOPIC alongside systems such as STAIRS, that also enable Boolean retrieval on strings in full-text documents. The docrep that consists of the set of words occurring in the document, and that is stored

[2]Not to be confused with the TOPIC that is currently being developed in Germany as a frame-driven IR system [Hahn,1987] that even may generate abstracts. This also is a very interesting development, but outside the scope of this paper.

in the inverted file, acts as a primary access mechanism. At retrieval time the original document is consulted to obtain information on the proximity of the words. Thus it might be said that the document representation so far consists of terms from the complete text of the documents, which makes it a typical derived index.

However, the second stage that is grafted on top of this retrieval engine, is a knowledge representation tool, that may be thought of as essentially a weighted thesaurus, but that is implemented as a rulebase. Concepts are arranged in trees, or rather acyclic graphs, in which the strings that should occur in the documents are the leaves. The occurrence of such strings, using Boolean and/or proximity operators, is taken as weighted proof for the relevance of the higher concept and such concepts in their turn support other concepts (see fig. 2). In a typical TOPIC-application many of such concepts (*topics*) will be built in advance by an information specialist, thus effectively adding a knowledge base to the system. Subsequently the user can build and add topics of his own and those topics may or may not be accessible to other users.

This second layer may be considered part of the document representation too. Indeed, if a topic is built and documents are recognized by the rules in that topic, these documents are added to a list with postings for that topic. Thus it is relatively easy to extract the sets of topics that may be said to belong to a document (i.e. score above a threshold for that document) and we have done this in order to conduct the experiments. We consider such topics also to be a document representation, one that really is separate from the inverted file and complete document mentioned above.

One might ask what would be the difference between the *topics* of TOPIC and the entries of an orthodox classification system or a thesaurus. The answer is that the topics here ultimately are defined as properties of documents in stead of in semantic terms. This gives the system a great flexibility, but also ample opportunity for snap decisions, ad hoc constructs and heuristics that may work fine in small collections, but break down when applied to big databases.

A possible reason for this is that in big databases different sub-populations of documents will come in existence, that all cover more or less an identical subject, but approach it from widely different angles and (therefore) will use different vocabularies. There is a distinct danger that the assessment of the weights in such cases will become progessively more subjective, whereas attempts to introduce objective methods (e.g. statistics) will in fact cause a return to the frequency-based models.

An extensive summing up of the shortcomings of TOPIC is given in [BasisPlus, 1990], but it should be noted that this report was written by an unsuccessful competitor for the library system of Tilburg University.

```
GENERAL-MOTORS
    * 1.00 GM-COMPANIES
           ** 0.50 GENERAL-MOTORS-ACCEPTA-PHRS Phrase
                  *** "general"
                  *** "motors"
                  *** "acceptance"
                  *** "corp"
           ** 0.50 "gmac"
           ** 0.50 "hughes aircraft co"
    * 0.50 GM-PEOPLE
           ** 1.00 GM-EX-CEO
                  *** 1.00 "roger smith"
           ** 1.00 GM-PRES
                  *** 1.00 LLOYD-REUSS Paragraph
                         **** "lloyd"
                         **** "reuss"
           ** 1.00 GM-CEO
                  *** 1.00 ROBERT-STEMPEL Paragraph
    * 0.50 GM-PRODUCTS
           ** 0.50 "pontiac"
           ** 0.50 "oldsmobile"
           ** 0.50 "buick"
           . . . . .
```

Fig. 2. *Topic* from TOPIC system (indent added).

CLARIT

The CLARIT system works on totally different principles. The most important component of the system is a dictionary builder, that tries to extract the most informative phrases from NL-texts[Evans, 1991], although more components, such as a retrieval interface are being added. The most interesting part is the said dictionary builder, that creates a list of NPs (noun phrases). The designers of the system call such a dictionary a *first order thesaurus* but the point should be stressed that no semantic relations are defined between the individual terms.

CLARIT's approach to indexing a document is as follows: The first activity consists of readying the document for processing by normalising the characterset etc. The document is then parsed by a very robust parser, that identifies the noun phrases (NP's) in the document and extracts them to a list of *candidate* NP's.

These NP's are scored by applying various frequency-based formulas to the individual words, in the course of which they are also compared with the characteristics of a domain corpus and a general English corpus. Word co-occurrence statistics are not considered.

```
# 140
\=
     2.322273 (nasdaq) () 0
     1.161136 (est) () 0
     0.387045 (toronto) () 0
     0.379069 (wholly owned subsidiary) () 0
     ........
\=
\>
     5.392150 (petroleum) () 0
     5.392150 (prnewswire) () 0
     0.928909 (asset) () 0
     0.928909 (subsidiary) () 0
     .........
\>
\?
     2.818624 (prnewswire giant) () 0
     2.696075 (prnewswire) () 0
     0.122549 (giant) () 0
     1.985585 (pacific petroleum incorporated) () 0
     ........
\?
```

Fig. 3 Result of CLARIT indexing.

Finally the candidate terms are matched with one or more lists of certified terms (a general English corpus and, if applicable, a domain corpus), thus creating three groups: *exact, general* and *novel* terms, based on the fact whether an exact match is found, the candidate NP consists of a constituent or sub-term of a certified term, or that the NP is a new term.

The end product of a CLARIT-indexing run is a document representation that consists of a weighted list of such terms, sorted out in exact, general and novel terms (fig. 3). Tests run on rather small document collections as publicised by Evans and others show exceptionally good performance as compared to human indexers [Evans, 1991]. However, these tests are limited to the comparison of individual documents and not, as we intend to do, taken over sets of documents.

First discussion.

As we already indicated, CLARIT is a typical derived indexing system in that the document representation is derived from the original document. One might argue that the knowledge in the parser is external knowledge, but this certainly is not the kind of knowledge into which we may map the documents or concepts occurring in the documents. The same is true for the various frequency-based formulas that are used by CLARIT. Such formulas capture the intuïtive notion that the frequency with which a word occurrs, is in some way meaningfull for its informative value, but they do not

contain the semantic or pragmatic knowledge that is commonly associated with *meaning* and *understanding*. For instance: the sentence

Dog, dog, dog, dog, dog, dog!

might well land this same document in the lap of somebody, who is looking for literature on dogs, because the word *dog* now possibly scores above some such statistic treshold. But this certainly does not mean that this document is about dogs!

On the other hand the certified list or lists against which the candidate phrases are matched, certainly is a knowledge representation, albeit a very shallow one.

The fact that terms exist that are composed of several words, thus improving precision, gives CLARIT a decidedly pre-coordinate flavour, although of course such terms also may be combined again at query time to create new concepts to search for.

Things are different if we consider TOPIC. The first docrep, which is used to access the document, is the set of all occurring word strings in the document, stored in the inverted file. Moreover, the user may restrict his use of TOPIC to those strings in various combinations with Boolean and proximity operators and in that case TOPIC does not essentially differ from older systems like STAIRS.

But the creation of topics as described above, means that a knowledge system comes into existence that is independent from the document representations. If a document 'fits' the rules for a certain topic, it may be considered as *assigned* to that topic, rather than that the topic is derived from the document. It might be argued that humans index a document in very much the same way: by forming hypotheses about the topicality of the document and by testing those hypotheses by checking for the existence of other concepts and strings in the document. TOPIC really does the same thing.

Also, as individual words and concepts are used to construct bigger concepts and these concepts are available at query time, TOPIC, too, resembles the older pre-coordinate systems. Contrary to CLARIT, these concepts are assigned to the documents, not derived from them.

This makes the philosophy behind both systems very dissimilar. TOPIC departs from concepts that are created by the user more or less independently from the database and which reflect his interests, but not necessarily the contents of the database. The system then tries to find evidence in terms of strings and combinations that the concept may be found in a document. Such knowledge is incremental and some claim that it may soon lead to conflicting *topics* (see the arguments put forward in [BasisPlus, 1990]). Also, there naturally is no support for topics and concepts, that never have been declared. This may result in poor performance, when a user approaches the system with a new need.

CLARIT, on the other hand, tries to identify such parts of the document as give a good indication what the document is about and creates an index of keywords and key phrases. This supposedly makes for a very regular performance as all documents are treated equal; but as we will see this is not necessarily the case. The statistical nature

of the frequency-based formulas applied by CLARIT carries an inherent danger of important terms narrowly missing some treshold.

So both system try to escape from the flat knowledge representation that is stored in the docreps of the inverted file of keywords type: CLARIT by extending the keywords to (weighted) key phrases, TOPIC by adding a user-supplied semantic hierarchy to the docreps. What the systems are actually attempting, is to try and improve on the older, single keyword systems in the face of the evidence brought forward by Salton and Cleverdon, as we will see presently.

That the performance of 'normal' free text databases of the STAIRS-type (with a complete inverted file of single keywords) was insufficient, was demonstrated by the Blair/Maron study. On the other hand, already in 1983 it was stated by Salton and McGill that "*...the simple uncontrolled indexing language produce the best retrieval performance, while the controlled vocabulary and phrases (simple concepts) furnished increasingly worse results.*" (Salton, 1983, p.102). They came to this conclusions after extensive testing of both automatic indexing (Salton and SMART) and manual indexing (The CRANFIELD experiments by Cleverdon). To quote the last author: "*...as narrower, broader or related terms are brought in ... performance decreases... The simple concept (phrase) index languages were overspecific.*" (in:Salton, 1983, p.102).

As Salton observed, these results are rather counterintuitive. It seemed that the adding of knowledge to the document representation actually lowered the performance of the whole system. Of course the SMART and CRANFIELD tests were designed to operate in a large and heterogeneous user population and the addition of thesaurus-like tools and multi-word concepts probably created an environment that was too specific to accommodate such a population. Still, those are not auspicious words under which to start the exploration of two systems that are either based on a thesaurus-like structure with broader and narrower-term relations, like TOPIC, or in the case of CLARIT, on the assumption that multi-word phrases are better index terms than single words.

The approach taken by TOPIC seems to counter the problems by placing the onus of building a thesaurus on the individual user, thus enabling him to concentrate on the relations he deems necessary and to ignore other relations. The designers of CLARIT concentrate on the possibilities offered by newer parsing algoritms and the probabilistic model to extract such phrases as are most descriptive for the document. It is interesting to note that the conlusions of the CRANFIELD-experiments of Cleverdon were based on manual indexing: so unless CLARIT outperforms humans on detecting the salient terms in a document, the observations of Cleverdon on the relatively bad performance of such phrase indexes seem to remain valid.

Company	: Giant Pacific Petroleum Inc. (GPPXF)
Industry	: Oil Drilling; Oilfield Equip & Services
Subject	: Acquisitions, Mergers, Takeovers
Market Sector	: Energy
Region	: Canada

Fig. 4 Original classification of document 140.

Methodology

Of course there is the question whether it is legitimate to compare two systems that differ so widely. But as they both claim to offer very similar services (the retrieval of full-text documents from a document base), there should be no reason why they should not be compared. The question is, which methods should be used.

We wanted to test how the two systems, or rather their docreps, would perform on a document base of free text documents on general subjects. In this section we will outline the methods used, taking the work of White and Griffiths as an example. While working on the TOPIC-system, we noticed a peculiarity in the demo database, which raised some new questions. We will cover these questions separately.

Selection and preparation of the databases

White and Griffiths used existing databases, each containing several millions of bibliographic records with attached descriptors. We had no such databases available for CLARIT and the TOPIC database of the library of Tilburg University contained no documents of the kind we wanted to use. Therefore we decided to use the demonstration database of 200 documents (500 Kb) that comes with TOPIC (eventually 128 documents were used). The database contained articles from Wall Street Journal. Thanks to the co-operation of Drs T. van den Aker of the computer centre of our University we were able to obtain a list of document-titles with the *topics* that were considered by the TOPIC-system to be relevant for each title. The topics themselves were also part of the demo database.

Initially we wanted to use the documents exactly as they occurred in the demonstration database. Then we found that to each of the original documents entries from a classification were attached (Fig. 4) and that this classification was also searched by the TOPIC system. We thought this a rather unfair tactic to use in a demonstration database and it made the records in their original form unsuitable for testing purposes so we decided to delete these classifications from the documents and to create a new TOPIC-database. The rather interesting question if and how much the classifications in the original database influenced the performance of the TOPIC system will be addressed elsewhere.

140 Acquisition of Patriot Energy Company Ltd. Announced

TOPIC	evidence	weight	CLARIT
TRADE-ACTION	0.55	5.392150	prnewswire
MERGER-ACTIVITY	0.83	5.392150	petroleum
MERGER-ACQUISITION	0.86	2.818624	prnewswire giant
FINANCIAL-TOPICS	0.43	2.696075	prnewswire
COMPUTER-PRODUCTS	0.80	2.322273	nasdaq
IBM-PRODUCTS	0.80	1.985585	pacific petroleum incorporated
IBM	0.48	1.861511	announced calgary
COMPUTER-COMPANIES	0.48	1.840358	calgary
		1.797383	petroleum
		1.283685	giant pacific'
		1.254242	vancouver stock exchange
		1.226905	vancouver
		1.161136	pacific'
		1.161136	est
		.....	

Fig. 5 Document representations in TOPIC and CLARIT.

It may be argued that using TOPIC's document base gives this system an unfair advantage over CLARIT. We don't see why, especially not after the excision of the classifications. Of course we could have chosen a different collection of documents, but then we would have had to construct a new hierarchy of topics and we wanted to avoid the criticism to have worked with a substandard set of topics.

We created for every record in the test collection a document representation to be used in qualitative comparisons, both for the CLARIT and for the TOPIC database (fig.5). Although both docreps consisted of weighted terms, we decided to ignore the weights in the quantitative considerations and disregarded the differences between the three term-groups of CLARIT. Also the very long CLARIT-lists generated for each document were truncated at a weight below 1.0000 or after thirty-five terms, whichever came first. Using this threshold, we found a total of appr. 3400 postings of 2000 different terms in the 128 records of the CLARIT-set: for TOPIC we noted 1215 postings of 103 TOPIC-terms.

It should be noted that the topics that constituted our TOPIC-docrep, are all non-terminal nodes of trees, whose leaves are literal strings. Every posting in the docrep signifies the occurrence of at least one lower topic or string and although both the higher and the lower topic are in our docrep, the string(s) that caused the firing of the topics, are not included. We extracted those strings separately (see below, *virtual docreps*) and found a total of 1351 postings of 205 distinct literals.

The next step was the creation of subsets of related documents in a manner that was independent of the indexing processes to be tested. The method followed by White and Griffiths, who used co-citing and co-referencing, could not be used for this database because of the nature of the documents. However, the same classification that we had to cut away from the original documents, suggested itself as an independent system (it was not created by TOPIC, we only omitted it from the tests because it was an enrichment of the original document). Thus we created a dictionary of the terms in the *industry-, subject* and *market*-fields of that classification and asked two independent persons (not professional indexers, but knowledgeable in the general field of the documents) to check documents and dictionary for consistency. Subsequently this classification was judged too unspecific to identify really interesting subsets. Then we asked them to identify six groups on the general subjects of networks, mergers, war/violence, software, legal matters and drugs/pharmaceuticals, and to identify in each group at least ten documents that were most alike. We took the intersection of these groups, thus obtaining one group of six documents (violence), three groups of seven documents (software, mergers and legal matters) and two of eight (pharmaceutics and networks).

Collecting the data.

Then we collected the docreps of every document in every group. Referring to fig. 8 (in appendix I), we see the results for every group for both CLARIT and TOPIC. The captions ***post.*** and ***terms*** indicate how many postings were scored on each group by the two systems and how many different terms were involved. As was expected, CLARIT shows higher scores than TOPIC does, but as the cut-off point for CLARIT was rather arbitrarily set on 1.0000 or 35 terms, one should not look too closely to the exact figures. Also, the CLARIT-output laboured under some irksome irregularities, such as the mis-interpreting of quotes. We edited all quotes out from the CLARIT docrep so that 'IBM was afterwards read as IBM. Spelling errors that occurred in the documents we left alone, because such errors were met too by the TOPIC system.

Another peculiarity of the CLARIT-output was, that the same phrase might occur more than one time in a single docrep; sometimes with varying and sometimes with equal weight As we were only interested in binary values (a term occurs in the docrep, or it does not occur), we disregarded double postings when computing the frequencies.

To get an insight in the performance of the systems in recognising similar documents, we had to find for each group the terms that spanned the groups totally or partially (i.e. that occurred in more than one docrep) on the assumption that such terms indicate similar properties. The exact figures may be found under the caption ***spanning*** for all terms spanning 1-8 records. Figure 6 gives an impression of the relative scores.

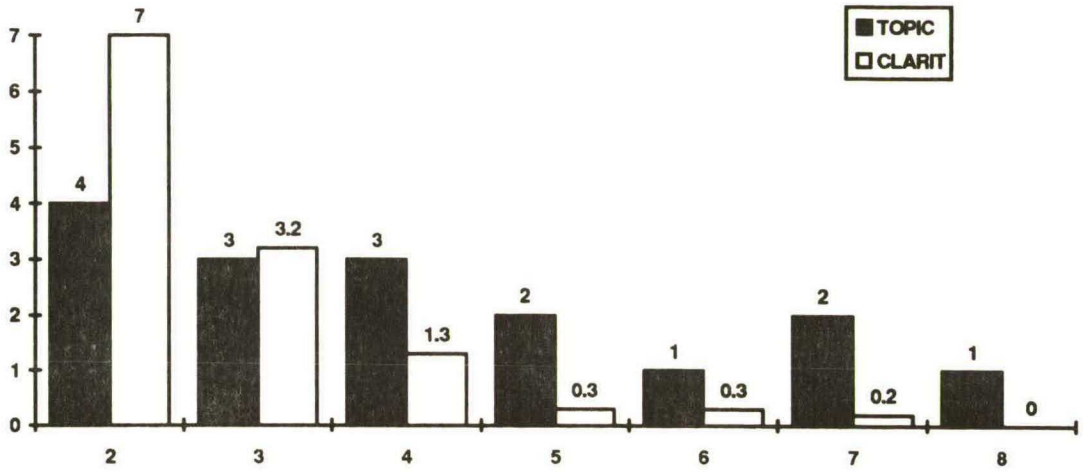

fig 6: average # terms spanning 2-8 docreps..

Then a second measure has to be found to check if those terms do not lump too many documents together. The ideal terms would span all records in the cluster and would occur in no other records. Therefore the number of documents in a cluster, that is spanned by a particular term is not complete as a measure without the frequency of that same term in the complete database. Therefore we added for each cluster a second and a third table under both TOPIC and CLARIT. In the second table the average frequency of the spanning terms in the database is shown.

In the TOPIC docrep occur topics like NUMBERS, POSITIVE-INDICATORS etc. that in itself have no bearing on the topicality of the document, but that are used to support other topics. Such topics typically have a very high frequency, whereas other topics might occur in the same spanning group that *do* have a bearing on the topicality of the document, and that may have a much lower frequency. Therefore we also included for each system a third table that shows the frequency of the terms with the lowest frequency. We have omitted the average and lower score of all terms where n<3 because terms that span less than 3 records in our groups were considered to be of no importance for identifying the group as a whole (fig. 7).

As we see, CLARIT displays far less spanning terms than TOPIC. Only in the software-group we find a CLARIT-term spanning the complete cluster and that term ('*software*') occurs in one of every five documents (19.6%).

TOPIC has eight terms that span all documents in three of the six clusters: the average score of the frequencies is 38.5%, the average lowest frequency of these terms is 32%. When we compute the scores taken over all terms spanning more than half of the documents in all clusters, we find 29 terms with an average frequency of 26.7 (19 for terms with lowest frequency) for TOPIC and five terms with avg. 7 and avg-lowest 6 for CLARIT.

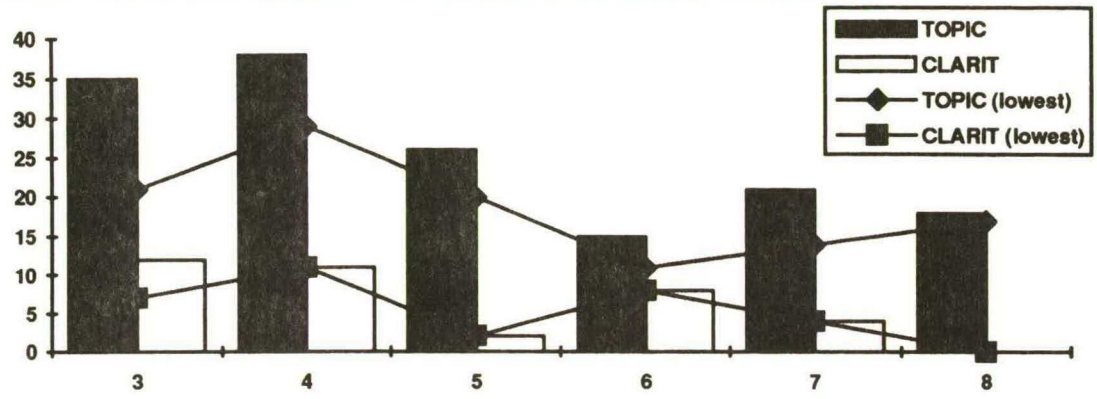

Fig. 7. Average and lowest frequency for terms spanning 3-8 docreps.

Plotting these values in a precision/recall graph, we obtain the two graphs of figure 8. The centre of the CLARIT-graph (i.e. the point where the number of points above and below, c.q. left and right are equal) lies at <0.2, 0.4>, the TOPIC centre at <0.08, 0.5>, confirming the general impression that TOPIC performs better in recall, whereas CLARIT is better in precision.

Virtual docreps.

Another interesting question is how high (or low) the agreement between the two systems lies (agreement understood as the number of identical terms in the docreps that two different indexing systems create for the same document). An obvious problem in comparing the TOPIC-docreps with the CLARIT-docreps was, that the TOPIC-docreps consisted of artificial terms, whereas the CLARIT terms were strings occurring in the documents - assigned vs. derived terms. The agreement as displayed in [Evans,1991] could therefore not be computed immediatly.

What we did was first expanding every TOPIC term to those strings in the original document on which the *topic* had fire*d*, but that are not visible in the TOPIC docrep as displayed in fig. 5. We will call these invisible docreps *virtual docreps*.

In the list with topics that came with the database, we found 495 strings that were considered by the authors as important. Checking in the 128 documents of the test, it was found that 205 of these leaves actually occurred. Comparing this list of 205 terms with the 2019 terms from the truncated CLARIT-list, we found that the number of terms that are both in the CLARIT docreps and in the TOPIC docreps was surprisingly small. Only twenty-six terms were identical. Of these 14 are proper names or acronyms. Checking how many times such TOPIC-leaves were *constituents* of CLARIT terms resulted in 424 cases (note that in this last case the CLARIT-term "computer company" would score on 'computer', on 'company' and on 'computer company' if all three existed as separate TOPIC-leaves).

This is rather surprising, because this result does not tally with the high level of agreement reported by Evans [Evans, 1991]. If we take TOPIC to be an alternative indexer that performed on the same level as the human indexers in the Evans-

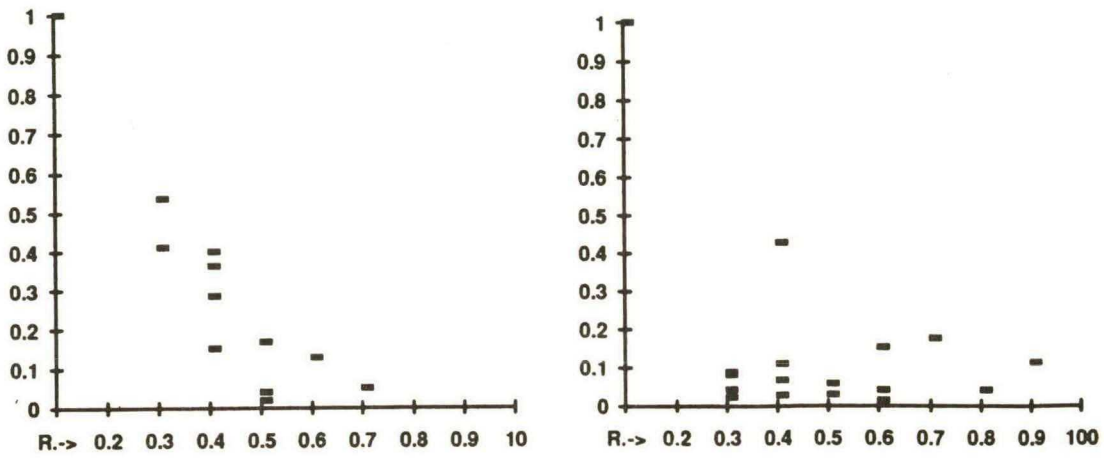

Fig 8. Precision vs. recall in CLARIT (a) and TOPIC (b)

experiment, one would expect on grounds of the tables on page 51 and 52 of the CLARIT evaluation that the agreement by words would be higer (typical 20-40) in every article and at least a few terms in every docrep would overlap. This evidently is not the case in our database: TOPIC and CLARIT seem to group its documents on widely different terms. Of course TOPIC is not a human indexer, but at least the keywords in the topics are selected by humans, so one would expect more agreement. Possibly this lack of agreement is a result of the assigned vs. derived technique, the assignments of TOPIC converging towards more general concepts, while the compound terms of CLARIT diverge towards ever more specific terms.

Conclusions and suggestions.

It should be stressed that the differences that the two systems display relative to each other in the quantitative tests, should not immediatly be translated in qualitative judgements. Offhand TOPIC seems to score better when groups of documents are to be retrieved, that cover a broad concept, or when a concept is described using many different, but identifiable terms. CLARIT seems to perform better when the documents display a marked terminology, because such terms are readily recognized against the background of the corpus.

This would lead to the conclusion that the TOPIC-system is more apt for big libraries that cover a rather wide spectrum of subjects. The very specificity that CLARIT displays, would point to a possible use in smaller document collections, or collections that limit themselves to a very specialized subject with users that know the specific terms of the trade.

Observing that the retrieval engine of TOPIC works with a derived dictionary, one might ask if CLARIT, itself a derived dictionary system, could be applied to create the TOPIC dictionary for subsequent use with the 'intelligent' topics.

There are some problems in this respect. The evidence that TOPIC looks for in order to decide on the assignment of the document to a topic, is the occurrence of any kind of string (not only NP's) or combination of strings (positional information included). CLARIT, on the other hand, only looks for NP's and discards all positional information after the candidate terms are generated. However, an inspection of the leaves of TOPIC hierarchies shows that they are almost exclusively nouns and noun phrases, so there is no inherent reason why they could not occurr in a CLARIT-dictionary of the document.

A second condition for inclusion would be that the frequency of the noun, or noun phrase, is such that it will be kept by the various statistics that CLARIT applies. This, however, seems to offer not much perspective as the agreement between CLARIT and TOPIC seems to be very low in this respect. Or, the words and phrases that CLARIT extracts, typically are not the words and phrases that have been included in the topics.

The various counts and experiments in this explorative paper should be followed up by more research in bigger databases, using a real-life test. After a solid measuring system for these and similar programs has been arrived at, the human-selected TOPIC leaves should be be replaced by phrases that have been selected by CLARIT (or from a CLARIT-generated list) in order to check if performance will improve.

Bibliography

Ajiferuke, I.;C.M. Chu:"Quality of indexing in library and information science databases" Online Review 13, 1: febr. 1989.

Appelbaum, L.A.; Tong, M. :"Conceptual information retrieval from full text" RIAO88

BasisPlus(?): "A Clarification of Claims made by Verity Corp. vis-a-vis the Capabilities of BASISplus" Information Dimensions, Inc. La Jolla, California, may 1990.

IBM:"Storage and Information Retrieval System/Virtual Storage (STAIRS/VS): General Information" IBM Program Product 5740-XR1 1976

Cleverdon, A.W. :"Optimizing convenient on-line access to bibliographic databases" Inf. Serv. Use 4 pp.37-47, (1984)

Cleverdon, C.W.; Keen, E.M. :"Aslib-Cranfield research project" Cranfield institute of technology, Cranfield, England 1966

Evans, D.:"The CLARIT project", Laboratory for computational linguistics, Carnegie Mellon University, 1991.

Evans, D.; Ginther-Webster, K.; Hart, M.; a.o.:" Automatic indexing using selective NLP and first-order Thesauri" RIAO91, p.624-643

Foskett, A.C.:"The subject approach to information" London 1969, 4th edition 1982

Fung, R.; Crawford, S.L.; Appelbaum, A.; Tong, R.M. :"An architecture for probabilistic concept-based information retrieval " SIGIR 1990 :979: p.455-467.

Hahn, U.; Reimer, U.:"Topic parsing: accounting for text macro structures in full text analysis". in: Lamersdorf (ed) "IFIP conference proceedings of 1987", North Holland, 1988.

McCune, B.P.; Tong, R.; Dean, J.:e.a.:"RUBRIC, a system for rule-based information retrieval" IEEE transactions on software engineering. 1985, p.939-944.

Paijmans, H.:" An inventory of models in Information Retrieval." in: Proceedings AIIR, Tilburg 1992, p. 6-18.in: Paijmans, H.; Weigand, H.:" Workshop on Artificial Intelligence and Information Retrieval: proceedings." Instituut voor Taal en Kennis Technologie (ITK) Tilburg 7 mei 1992

Salton, G. ;M.J. McGill :"Introduction to Modern Information Retrieval " New York [etc.] : McGraw-Hill, 1983. - 448 pp.

White, H.D.; Griffith, B.C.:"Quality of indexing in online data bases" Information processing and management, V.23, p.211-224, 1987

APPENDIX

Example of the documents in the database. Most figures have been drawn from this document.

Document 140.

ACQUISITION OF PATRIOT ENERGY COMPANY LTD. ANNOUNCED

CALGARY, Alberta, Feb. 12 /PRNewswire/ -- Giant Pacific Petroleum Inc. ('Giant Pacific') (NASDAQ, Vancouver: GPP) today announced that it has acquired all of the 5,333,528 common shares of Patriot Energy Company Limited ('Patriot') representing approximately 94 percent of the outstanding shares of Patriot, tendered pursuant to the take over bid offer for all the shares of Patriot.
Payment for the shares tendered will be made as soon as possible. The warrants to be issued pursuant to the offer will be listed for trading on the Vancouver Stock Exchange ('VSE') in due course provided that the listing requirements of the VSE are complied with. Giant Pacific intends to exercise its statutory rights under the compulsory acquisition provisions of the Business Corporations Act (Alberta) to acquire the remaining shares of Patriot and consequently to hold Patriot as a wholly owned subsidiary.

Giant Pacific trades on the VSE and NASDAQ. Its principal assets are oil and gas producing properties in Texas. Patriot is an active junior oil and gas exploration company which has holdings in western Canada with oil production in Saskatchewan.

The Vancouver Stock Exchange has neither approved or disapproved of the information contained herein.

/CONTACT: Bruce Weaver, president, Giant Pacific Petroleum Inc., Toronto, 416-941-9440/
10:35 EST

TOPIC	postings	terms	1	2	3	4	5	6	7	8	
violence	46	23	12	5	3	0	5	0	-	-	
software	129	44	20	2	5	8	1	3	5	-	
pharma.	32	22	18	1	1	1	1	0	0	0	
networks	127	42	20	5	1	4	4	1	5	2	
mergers	61	26	8	11	3	1	0	2	1	-	
legal	45	27	19	1	4	3	0	0	0	-	
	73	31	16	4	3	3	2	1	2	1	
	TOPIC: Average frequencies in %										
violence					44	0	30	0	-	-	
software					7	16	33	54	36.6	-	
pharma.					56	73	70	0	0	0	Spanning max.
networks					44	37	24.7	7.8	21.9	35	38.5
mergers					23.3	70	0	27.5	44	-	
legal					36.7	34.9	0	0	0	-	Avg.freq.
					35	38	26	15	21	18	26.7
	TOPIC: Lowest-frequency terms in %										
violence					4.7	0	10	0	-	-	
software					3.9	6	33	37	18.8	-	
pharma.					56	73	70	0	0	0	Spanning max
networks					44	8	6.2	7.8	9.5	33	31.9
mergers					13	70	0	23	44	-	
legal					7	14.7	0	0	0	-	Avg. lowest
					21	29	20	11	14	17	19.1
CLARIT	postings	terms	1	2	3	4	5	6	7	8	
violence	104	95	89	3	2	0	0	0	-	-	
software	199	170	154	10	3	1	1	0	1	-	
pharma.	179	150	134	10	3	1	1	1	0	0	
networks	223	191	169	10	6	5	0	1	0	0	
mergers	138	125	115	5	4	1	0	0	0	-	
legal	118	114	111	2	1	0	0	0	0	-	
	160	141	129	7	3.2	1.3	0.3	0.3	0.2	0	
	CLARIT: Average frequencies in %										
violence					19.7	0	0	0	-	-	
software					5	12	6.2	0	19.6	-	
pharma.					7.3	25	7.8	28.3	0	0	
networks					5.6	5.9	0	19.6	0	0	
mergers					16.5	25	0	0	0	-	
legal					14.9	0	0	0	0	-	
					12	11	2	8	4	0	7
	CLARIT: Lowest-frequency terms in %										
violence					14.7	0	0	0	-	-	
software					3.9	12	6.2	0	19.6	-	
pharma.					3.9	25	7.8	28.3	0	0	
networks					2.3	3.1	0	19.6	0	0	
mergers					3.9	25	0	0	0	-	
legal					14.9	0	0	0	0	-	
					7	11	2	8	4	0	6.1

ITK: P.O. BOX 90153 5000 LE TILBURG THE NETHERLANDS